Detect & Reject for Transferability of Black-box Adversarial Attacks Against Network Intrusion Detection Systems

Islam Debicha[1,2]✉, Thibault Debatty[1], Jean-Michel Dricot[2], Wim Mees[1], and Tayeb Kenaza[3]

[1] Royal Military Academy, Rue Hobbema 8, 1000 Brussels, Belgium,
debichasislam@gmail.com, {thibault.debatty,wim.mees}@rma.ac.be
[2] Université libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050 Brussels, Belgium
jean-michel.dricot@ulb.be
[3] École Militaire Polytechnique, Bordj El-Bahri 17, 16111 Algiers, Algeria
ken.tayeb@gmail.com

Abstract. In the last decade, the use of Machine Learning techniques in anomaly-based intrusion detection systems has seen much success. However, recent studies have shown that Machine learning in general and deep learning specifically are vulnerable to adversarial attacks where the attacker attempts to fool models by supplying deceptive input. Research in computer vision, where this vulnerability was first discovered, has shown that adversarial images designed to fool a specific model can deceive other machine learning models. In this paper, we investigate the transferability of adversarial network traffic against multiple machine learning-based intrusion detection systems. Furthermore, we analyze the robustness of the ensemble intrusion detection system, which is notorious for its better accuracy compared to a single model, against the transferability of adversarial attacks. Finally, we examine Detect & Reject as a defensive mechanism to limit the effect of the transferability property of adversarial network traffic against machine learning-based intrusion detection systems.

Keywords: Intrusion detection, Machine learning, Adversarial attacks, Transferability, Black-box settings

1 Introduction

The computer and networking industry is becoming increasingly important due to their growing use in various fields, which in turn leads to an escalation in the occurrence of cyberattacks. As a result, security is becoming a key concern of any network architecture and an active research topic. Intrusion Detection Systems (IDS) are one of the solutions presented to enhance network security by analyzing the traffic to identify any suspicious activity.

In order to detect intrusions, there are mainly two approaches: the first one is based on comparing traffic with a list of all known attack patterns, also called signature-based intrusion detection. Intuitively, this method gives excellent accuracy when dealing with

known attacks, however, this type of detector is incapable of detecting zero-day attacks, which is essentially what motivates the use of the second type of intrusion detection techniques, called anomaly-based intrusion detection. The latter approach relies on modeling the normal behavior of network traffic and later examining new traffic against this baseline.

Anomaly-based intrusion detection has been extensively studied, with most research using Machine Learning (ML) techniques to create a trustworthy model of activity due to their high accuracy, including deep learning which is considered a state-of-the-art technique in this field. However, recent studies have shown that machine learning in general, and deep learning in particular, are vulnerable to adversarial attacks where the attacker seeks to fool the models by inserting slight but specially crafted distortions into the original input [1].

Research in the field of computer vision, where this vulnerability was first discovered, has shown that adversarial images designed to fool a specific model can, to some extent, fool other machine learning models [2]. This is known as the transferability property of adversarial attacks. By exploiting this property, an attacker can build a surrogate intrusion detection system, create adversarial traffic for that detector, and then attack another intrusion detection system without even knowing the internal architecture of that detector, leading to a black-box attack.

To avoid this kind of vulnerability, we are conducting this research and the following are our contributions in this paper:

- To the best of our knowledge, this is the first study to examine the transferability of adversarial network traffic between multiple anomaly-based intrusion detection systems with different machine learning techniques in black-box settings.
- In addition, we construct an ensemble intrusion detection system to examine its robustness against the transferability property of adversarial attacks compared to single detectors.
- Finally, we investigate the effectiveness of the Detect & Reject method as a defensive mechanism to mitigate the effect of the transferability property of adversarial network traffic against machine learning-based intrusion detection systems.

2 Background

2.1 Related Work

Studies have shown that incorporating machine learning techniques can help improving the performance of intrusion detection systems. Aslahi-Shahri et al. [3] have proposed and explained the implementation of an intrusion detection system based on a hybrid support vector machine (SVM) and genetic algorithm (GA) method. In [4], the authors have proposed a novel evidential IDS based on Dempster-Shafer theory to take into account source reliability. Vinayakumar et al. [5] have proposed a highly scalable intrusion detection framework using deep neural networks (DNNs) after a comprehensive evaluation of their performance against classical machine learning classifiers. Alamiedy et al.[6] proposed an improved anomaly-based IDS model based on a multi-objective gray wolf optimization (GWO) algorithm, in which GWO is used as a feature selection

technique. Ghanem et al.[7] have proposed a cyber-intrusion detecting system classification with MLP trained by a hybrid metaheuristic algorithm and feature selection based on multi-objective wrapper method. In [8], the authors proposed a new binary classification model for intrusion detection, based on hybridization of Artificial Bee Colony algorithm (ABC) and Dragonfly algorithm (DA) for training an artificial neural network (ANN) in order to increase the classification accuracy rate for malicious and non-malicious traffic in networks. Nevertheless, we noticed that little or no attention was paid to the effect of adversarial attacks when proposing these solutions.

Szegedy et al.[9] was the first work to report the vulnerability of DNN to adversarial samples where they introduced imperceptible adversarial perturbations to handwritten digits images and succeeded to fool the DNN model with high confidence. This discovery has prompted a number of studies in the computer vision community, where several attacks and defenses have been proposed [10, 11, 12]. There are some works [2, 13, 14] dealing with the transferability of adversarial attacks. However, these studies were specifically designed for the field of image classification, where this particular vulnerability was first detected.

In recent studies, the effect of adversarial attacks on intrusion detection systems has been investigated. Wang [15] Inspected the performance of state-of-the-art attack algorithms against deep learning-based intrusion detection on the NSL-KDD dataset. Pawlicki et al.[16] evaluated the possibility of deteriorating the performance of a well-optimized intrusion detection algorithm at test time by generating adversarial attacks and then offers a way to detect those attacks.

Through our literature review, we did not find any studies on the transferability of adversarial network traffic between multiple anomaly-based intrusion detection systems with different machine learning techniques in black-box settings. Hence, we propose to conduct this study in order to fill this gap. Notice that in our recent work [17], we investigated the effectiveness of adversarial training as a defense for intrusion detection systems against these attacks.

2.2 Adversarial Attacks

Despite their considerable success in achieving high accuracy, machine learning algorithms in general and deep learning, in particular, have proven vulnerable to adversarial attacks, where crafting an instance with small intentional perturbations can lead a machine learning model to make an erroneous prediction [9].

The idea of generating adversarial examples is quite intuitive. It can be seen as the inverse process of gradient descent where, for a given input x and its label y, one tries to find model parameters θ that maximize the accuracy of the model by minimizing the loss function J. On the other hand, the adversarial examples are generated in order to minimize the accuracy of the model. Given the parameters θ, the loss function J is differentiated with respect to the input data x so as to find an instance x', as close to x as possible, that maximizes the loss function J.

One of the earliest and most popular adversarial attacks is called the Fast Gradient Sign Method (FGSM) [10]. This attack uses a factor ϵ to limit the amount of distortion in the original instance such that $\|x' - x\| < \epsilon$. One can think of ϵ as the attack strength

or the size of the introduced perturbation. An adversarial instance x' is devised like following:

$$x' = x + \epsilon \nabla J_x(x, y, \theta) \tag{1}$$

Projected Gradient Descent (PGD) [11] is another adversarial attack and is basically an iterative extension of FGSM applying the attack repeatedly. However, PGD initializes the instance, at each iteration, to a random point in the ϵ-ball around the original input.

Adversarial attacks can be classified into two types based on the attacker's knowledge of the attacked system: White-box setting where the attacker has full knowledge of the internal architecture of the attacked system. Black-box setting where the attacker has no knowledge of the internal architecture of the attacked system. In this paper, we create adversarial network traffic records against a DNN-based IDS in a white-box setting, and then the rest of the experiments are conducted in a black-box setting where we use these adversarial instances to attack the other ML-based IDSs without having access to their internal architecture or training data.

3 Proposed Approach

Previous work has shown that the accuracy of a DNN-based IDS can be significantly reduced when exposed to adversarial attacks [15, 16, 17]. In this section, we construct a DNN-based IDS and five other ML-based IDSs to examine whether the same adversarial instances designed for a DNN-based IDS can be transferred to other ML-based IDSs, which are trained on a different data set, without knowing anything about their internal architectures. We also build an ensemble intrusion detection by clustering the five ML-based IDSs to study whether having multiple classifiers voting prediction can be a defense against the transferability property of adversarial attacks. Finally, we implement the Detect & Reject method as a defense mechanism for the intrusion detection system and evaluate its robustness to transferable adversarial examples.

3.1 NSL-KDD Dataset

Many research papers in the intrusion detection community have used the NSL-KDD dataset to demonstrate the performance of their proposed approaches and this is very useful as it creates a common basis for comparison between these approaches. Published in 2009 by [18], this dataset is an improvement of the well-known KDD-CUP'99 which has a fundamental problem of record redundancy which leads to the bias of classifiers towards frequent records. This problem has been solved in the NSL-KDD dataset by proposing a balanced version of KDD-CUP'99 by removing the redundancy, thus providing a more accurate comparative analysis of the performance of different proposed intrusion detection frameworks.

The dataset contains 41 network traffic characteristics, covering three aspects: basic characteristics, content characteristics, and traffic characteristics. Many attacks are covered in this dataset; they can be further classified into four families of attacks: denial-of-service (DoS) attacks, probe attacks (Probe), root-to-local (R2L) attacks, and user-to-root (U2R) attacks. We use KDDTrain+ in our experiments by dividing it into 80%

and 20% for training and test data respectively. The training data is divided into two almost equal parts A and B to train the DNN-based IDS separately from other ML-based IDSs so as to examine the transferability property. The data used in the experimental part are summarized in Table 1 and their partitioning is illustrated in Figure 1 .

Table 1. Summary of the network traffic dataset.

	Normal	**DoS**	**Probe**	**R2L**	**U2R**
Training data A	26938	18371	4663	398	21
Training data B	26937	18371	4662	398	21
Test data	13468	9185	2331	199	10

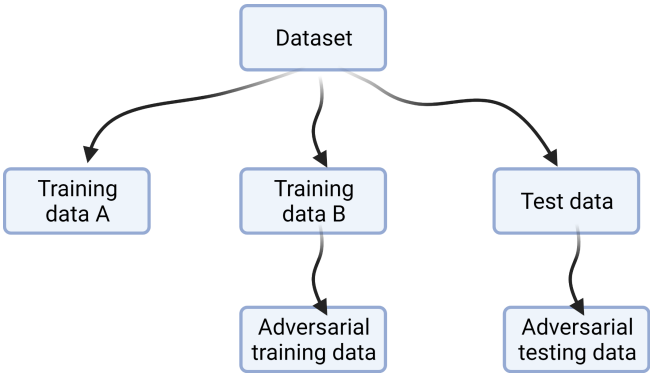


Fig. 1. The partitioning of the dataset for training and testing of IDS

3.2 Preprocessing

The network traffic included in this dataset is heterogeneous and contains both numerical and categorical values. Many machine learning algorithms do not support categorical values, hence the need for the numericalization step that transforms these categorical inputs into numerical values. In the case of the NSL-KDD dataset, the categorical features are ”flag”, ”protocol type” and ”service”. Another important aspect is feature scaling, which consists of converting all features to the same scale to ensure that all features contribute equally to the result and also to help the gradient-based ML algorithms converge faster to the minima. We restrict our study to a binary classification where we consider any type of attack as ”intrusion” and the rest as ”normal” traffic.

3.3 Building Anomaly-based Intrusion Detection Systems

TensorFlow [19] is used to build the DNN-based IDS, it consists of two hidden layers with 512 units each. As an activation function, we use Rectified Linear Unit (ReLU) to

increase the non-linearity. To prevent overfitting, a dropout layer with a 20% dropout rate is placed after each hidden layer. ADAM and categorical cross-entropy are used as an optimization algorithm and loss function respectively. In the end, the logits are converted to probabilities using a softmax layer. The final prediction is assigned to the highest probability class.

We acknowledge the use of Scikit-learn [20] to build five ML-based IDSs. The default settings were maintained. These five ML algorithms were selected due to their popularity in the ML community: Support Vector Machines (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and Linear Discriminant Analysis (LDA). We also construct an ensemble IDS by grouping these five ML algorithms where the final prediction is made using the majority voting rule.

3.4 Transferability of Adversarial Attack in Black-box Settings

In order to test the transferability property of adversarial attacks, we build a DNN-based IDS where we generate adversarial network traffic records in a white-box setting and then test them against five different ML-based IDSs. Note that the five ML-based IDSs are trained on a different dataset (Training data B) and the adversary records were generated without assuming any knowledge of the internal architecture of these five ML-based IDSs, which means that we are working under a black-box setting assumption.

Two adversarial attacks were implemented to generate adversarial network traffic records: FGSM and PDG. For this, we use Adversarial Robustness Toolbox (ART) [21]. The experiments are repeated by increasing the attack strength ϵ to investigate the amount of perturbation required for the adversarial attack to be transferred from the DNN-based IDS to the other five ML-based IDSs in black-box settings.

3.5 Defenses against the Transferability of Adversarial Attacks

Since the ensemble technique is known to increase accuracy over a single classifier [22], we want to examine whether it can also increase its robustness against the transferability of adversarial attacks in a black-box setting. To do so, we construct an ensemble IDS based on the previous five ML-based IDSs and use the majority voting rule to obtain the final decision.

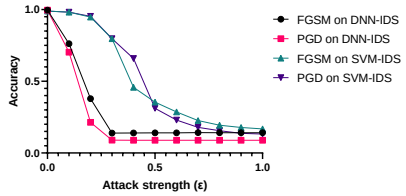
The second defense we consider is the Detect & Reject method [12], which involves training our IDSs to detect not only ”abnormal” and ”normal” traffic, but also a third class called ”adversarial”. Thus, whenever the IDS decides that a network traffic record is adversarial, it is rejected. We implement this method on the five ML-based IDSs and examine their robustness to the adversarial attack transferability property.

4 Experimental Results

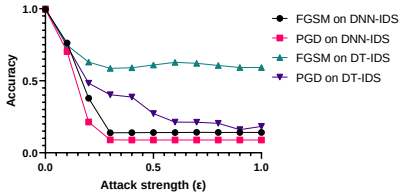
In this section, we present the results of experimenting our appraoch. Subsection (A) illustrates the effect of the transferability property of adversarial attacks on the five ML-based IDSs. In subsection (B), we examine the robustness of the ensemble IDS

against these attacks in black-box settings. Subsection (C) illustrates the robustness improvement of all IDSs after adding the detection and rejection mechanism to the five ML-based IDSs.

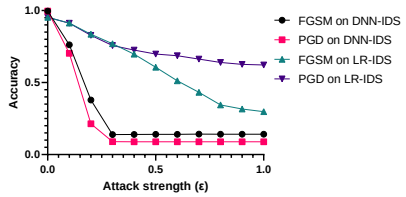
4.1 Transferability of Adversarial Attacks in Black-box Settings



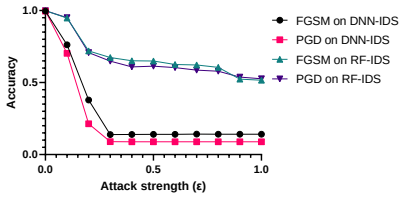
(a) Transferability from DNN-based IDS to SVM-based IDS



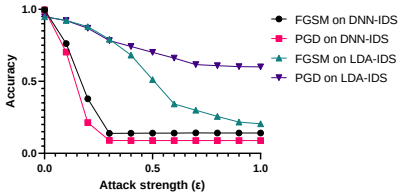
(b) Transferability from DNN-based IDS to DT-based IDS



(c) Transferability from DNN-based IDS to LR-based IDS



(d) Transferability from DNN-based IDS to RF-based IDS



(e) Transferability from DNN-based IDS to LDA-based IDS

Fig. 2. Transferability of adversarial attacks against ML-based intrusion detection systems in black-box settings

In this study, we use two adversarial attacks: FGSM and PGD to generate adversarial network traffic records from ”Test data”. These adversarial records are specifically designed to fool DNN-based IDSs since both attacks have access to the internal architecture of DNNs. As mentioned earlier, we train the DNN-based IDS using ”Training data A”, while the other 5 ML-based IDSs are trained using ”Training data B”. Adversarial traffic records are used at test time to attempt to mislead the IDSs. As shown in

Figure 2, increasing the attack strength (ϵ) further degrades the accuracy of the DNN-based IDS. When testing these adversarial network traffic records on the five ML-based IDSs, we find that their accuracy decreases, even though the attacks do not have access to their internal architectures. We also note that although the accuracy of the ML-based IDSs did not deteriorate as much as the DNN-based IDSs, some models were more vulnerable than others. This may be due to their differentiability property, i.e., they are composed of differentiable elements, since the decision tree and the random forest, whose accuracies were least affected, are non-differentiable models that are not amenable to gradient descent due to their Boolean nature, unlike SVM or logistic regression for example.

4.2 Ensemble Intrusion Detection System Robustness

Since the ensemble technique is known to improve accuracy over a single model, we investigate whether it could also improve robustness. To this end, we construct an ensemble IDS based on the five ML-based IDSs using the majority voting rule. The same setup as in the previous experiment is maintained, which means that the 5 ML models used to build the ensemble model are trained using the "Training data B". The adversarial traffic records are generated from the "Test data" using the FGSM and PGD attacks. These adversarial records are designed to fool DNN-based IDS since both attacks can only access the internal architecture of the DNN model. As shown in Figure 3, the ensemble IDS is not able to resist the transferability property of adversarial attacks, even though no information about the ensemble IDS was used to generate these adversarial records. This shows the ease of an evasion attack against an intrusion detection system without even knowing its internal architecture, simply by building a surrogate IDS (DNN-based IDS in our case) and generating adversarial network traffic for this surrogate model.

In practice, if we consider malicious network traffic, such as HTTP traffic that seeks to connect to dangerous URLs, like command-and-control [C&C] servers, an attacker could use adversarial attack techniques to disguise this malicious network traffic as normal traffic for the intrusion detection system while retaining its malicious aspects. This can be done by adding small amounts of specially designed data to that network traffic as a padding for example. Therefore, the attacker could evade the intrusion detection system.

4.3 Detect & Reject for Adversarial Network Traffic

In order to limit the effect of adversarial attacks in a black-box context, we implement the Detect & Reject method in each of the five ML-based IDSs. This method consists of re-training the model to detect not only "abnormal" and "normal" traffic but also "adversarial" traffic. PGD is used against "Training data B" to generate "Adversarial data". After that, the ML model uses a combination of "Training data B" and "Adversarial data" during the training phase to learn to distinguish the three classes. During the prediction phase, any network traffic record recognised as "adversarial" will be rejected. As shown in Figure 4, all five ML-based IDSs have improved their robustness against adversarial attacks. Decision Tree and Random Forrest, which is an ensemble

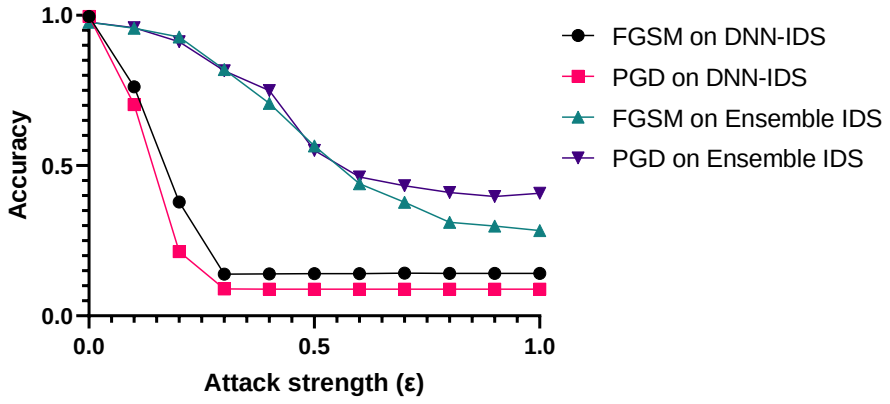


Fig. 3. Adversarial attack transferability from DNN-based IDS to Ensemble IDS

version of Decision Tree, have the highest detection rates of adversarial network traffic compared to the other IDSs.

5 Conclusion and Future Work

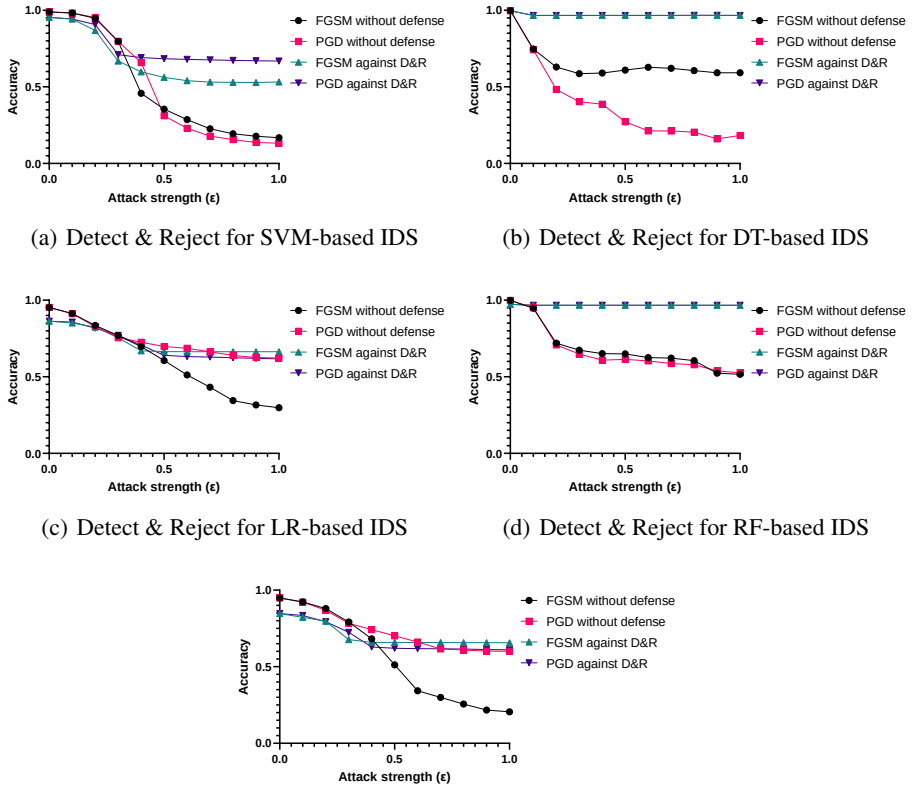
From an intrusion detection system perspective, adversarial attacks are a serious threat, as a small intentional perturbation of network traffic can mislead the system. To generate these adversarial records, the attacker must have access to the internal architecture of the machine learning model. However, by exploiting the transferability property of adversarial attacks, he can mislead other intrusion detection systems without having any knowledge about them. Ensemble IDSs, although known to improve model accuracy, are vulnerable to these attacks and thus cannot improve model robustness. On the other hand, Detect & Reject has shown through our experiments to be a suitable built-in defense for intrusion detection systems against adversarial attacks. An interesting future work would be to design more effective defenses to limit the effect of adversarial attacks against intrusion detection systems.

References

[1] Naveed Akhtar and Ajmal Mian. “Threat of adversarial attacks on deep learning in computer vision: A survey”. In: *IEEE Access* 6 (2018), pp. 14410–14430.

[2] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples”. In: *arXiv preprint arXiv:1605.07277* (2016).

[3] BM Aslahi-Shahri et al. “A hybrid method consisting of GA and SVM for intrusion detection system”. In: *Neural computing and applications* 27.6 (2016), pp. 1669–1676.



(a) Detect & Reject for SVM-based IDS

(b) Detect & Reject for DT-based IDS

(c) Detect & Reject for LR-based IDS

(d) Detect & Reject for RF-based IDS

(e) Detect & Reject for LDA-based IDS

Fig. 4. Detect & Reject as a defense against the transferability property of adversarial network traffic

[4] Islam Debicha et al. "Efficient Intrusion Detection Using Evidence Theory". In: *INTERNET 2020 : The Twelfth International Conference on Evolving Internet*. 2020, pp. 28–32.

[5] Ravi Vinayakumar et al. "Deep learning approach for intelligent intrusion detection system". In: *IEEE Access* 7 (2019), pp. 41525–41550.

[6] Taief Alaa Alamiedy et al. "Anomaly-based intrusion detection system using multi-objective grey wolf optimisation algorithm". In: *Journal of Ambient Intelligence and Humanized Computing* (2019), pp. 1–22.

[7] Waheed Ali HM Ghanem et al. "Metaheuristic Based IDS Using Multi-objective Wrapper Feature Selection and Neural Network Classification". In: *International Conference on Advances in Cyber Security*. Springer. 2020, pp. 384–401.

[8] Waheed Ali HM Ghanem et al. "An efficient intrusion detection model based on hybridization of artificial bee colony and dragonfly algorithms for training multilayer perceptrons". In: *IEEE Access* 8 (2020), pp. 130452–130475.

[9] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).
[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).
[11] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).
[12] Kathrin Grosse et al. "On the (statistical) detection of adversarial examples". In: *arXiv preprint arXiv:1702.06280* (2017).
[13] Cihang Xie et al. "Improving transferability of adversarial examples with input diversity". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2730–2739.
[14] Yantao Lu et al. "Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 940–949.
[15] Zheng Wang. "Deep learning-based intrusion detection with adversaries". In: *IEEE Access* 6 (2018), pp. 38367–38384.
[16] Marek Pawlicki, Michał Choraś, and Rafał Kozik. "Defending network intrusion detection systems against adversarial evasion attacks". In: *Future Generation Computer Systems* 110 (2020), pp. 148–154.
[17] Islam Debicha et al. "Adversarial Training for Deep Learning-based Intrusion Detection Systems". In: *ICONS 2021 : The Sixteenth International Conference on Systems*. 2021.
[18] Mahbod Tavallaee et al. "A detailed analysis of the KDD CUP 99 data set". In: *2009 IEEE symposium on computational intelligence for security and defense applications*. IEEE. 2009, pp. 1–6.
[19] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.
[20] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
[21] Maria-Irina Nicolae et al. "Adversarial Robustness Toolbox v1. 0.0". In: *arXiv preprint arXiv:1807.01069* (2018).
[22] Lior Rokach. "Ensemble-based classifiers". In: *Artificial intelligence review* 33.1 (2010), pp. 1–39.